Carcinogenesis vol.31 no.1 pp.111–120, 2010
doi:10.1093/carcin/bgp273
Advance Access publication November 11, 2009

Genome-wide association studies in cancer—current and future directions

Charles C.Chung[1], Wagner C.S.Magalhaes[1,2], Jesus Gonzalez-Bosquet[1] and Stephen J.Chanock[1,*]

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, 20892-4608, USA and [2]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CEP 31270-910, Belo Horizonte, MG, Brazil

*To whom correspondence should be addressed. Tel: +1 301 435 7559; Fax: +1 301 402 3134;
Email: chanocks@mail.nih.gov

Genome-wide association studies (GWAS) have emerged as an important tool for discovering regions of the genome that harbor genetic variants that confer risk for different types of cancers. The success of GWAS in the last 3 years is due to the convergence of new technologies that can genotype hundreds of thousands of single-nucleotide polymorphism markers together with comprehensive annotation of genetic variation. This approach has provided the opportunity to scan across the genome in a sufficiently large set of cases and controls without a set of prior hypotheses in search of susceptibility alleles with low effect sizes. Generally, the susceptibility alleles discovered thus far are common, namely, with a frequency in one or more population of >10% and each allele confers a small contribution to the overall risk for the disease. For nearly all regions conclusively identified by GWAS, the per allele effect sizes estimated are <1.3. Consequently, the findings of GWAS underscore the complex nature of cancer and have focused attention on a subset of the genetic variants that comprise the genomic architecture of each type of cancer, which already can differ substantially by the number of regions associated with specific types of cancer. For instance, in prostate cancer, there could be >30 distinct regions harboring common susceptibility alleles identified by GWAS, whereas in lung cancer, a disease strongly driven by exposure to tobacco products, so far, only three regions have been conclusively established. To date, >85 regions have been conclusively associated in over a dozen different cancers, yet no more than five regions have been associated with more than one distinct cancer type. GWAS are an important discovery tool that require extensive follow-up to map each region, investigate the biological mechanism underpinning the association and eventually test the optimal markers for assessing risk for a disease or its outcome, such as in pharmacogenomics, the study of the effect of genetic variation on pharmacological interventions. The success of GWAS has opened new horizons for exploration and highlighted the complex genomic architecture of disease susceptibility.

Introduction

The history of human genetics has focused on mapping regions of the genome that can explain part or all of a disease or human trait. With the generation of a draft of the human genome in 2001, geneticists quickly set out to comprehensively annotate the genome and apply the evolving knowledge of the pattern of genetic variation to investigate both monogenic, Mendelian disorders and complex diseases, the latter of which by nature are polygenic (1–4). Until recently, the scope and breath of human variation was certainly underappreciated until the advent of early maps of common variants, such as the single-nucleotide polymorphism (SNP), the most common variant in the genome (1,5–7). It is notable that a comprehensive set of genetic variation has shifted the analysis paradigm to finding genetic contributions to complex disease, whereas the capacity to capture environmental exposures and lifestyle decisions is far more rudimentary, even though these factors are essential for understanding complex diseases and traits.

For many years, human genetics has successfully mapped uncommon mutations with large effect sizes in studies conducted in families or special populations, such as the *BRCA1/BRCA2* mutations in Ashkenazi women with breast cancer and ovarian cancer (8). The search for highly penetrant mutations in familial aggregation has been based on genetic linkage analysis, an approach that has used microsatellite markers across the genome to scan for markers that segregate within a family (9,10). Based on the identification of linkage peaks using rigorous statistical approaches, follow-up of regions was pursued based on strong signals. Because of the wide spacing of markers across the genome, signals often pointed to regions over multiple megabases that in turn required sequencing large regions of the genome in search of the causative mutations, a daunting task in scope and until recently hampered by technical limitations. Nonetheless, successes in families loaded with melanoma, breast cancer and sets of cancers (Li-Fraumeni Syndrome) (8,11–14) are notable and provided an important substantiation of the approach of using markers indirectly. In retrospect, the use of markers to conclusively identify regions for detailed analysis has been an important lesson for mapping germ line genetic variants associated with risk for cancer, but the approach yielded only mutations with very strong effects.

Over the past 20 years, a parallel approach has been pursued to discover common genetic variants that confer susceptibility to different types of cancers. Initially, association studies were conducted using a handful of annotated genetic variants for which a strong hypothesis could be formulated. In a genetic association study, the analysis consists of a comparison of the distribution of a marker allele between cases and controls, in search of a statistical difference that can be reflected in an estimated effect size—usually quite small compared with mapped linkage signals due to highly penetrant mutations. Naively, at first, investigators searched for alleles with high estimated effect sizes (e.g. per allele odds ratios > 2.0), but with time, it has become apparent that common alleles confer small risk overall in sufficiently large case–control studies of unrelated subjects, the primary study design for association analyses (15).

Nominally, investigators focused on SNPs that altered the coding sequence and resulted in a non-synonymous change, namely a shift in the amino acid sequence of the protein. The approach was predicated on a more simplistic model: changes in the amino acid content would lead to a pronounced (e.g. measurable) change in function and thus influence the disease or trait of interest. Due to the inadequately sized studies, issues of study design and the overestimation of effect size, nearly all published candidate gene association studies, probably represent false positives. In this regard, the candidate gene approach has yielded very few notable findings, namely those that are conclusive and do not represent false positives. To date, perhaps a handful have been adequately replicated and confirmed in follow-up studies. For example, *GSTM1* null and *NAT2* slow acetylator genotypes have been associated with increased overall risk of bladder cancer and could account for up to 31% of the disease because of their high prevalence (16). Similarly, candidate genes have shown robust findings for a promoter SNP in *TNF* in non-Hodgkin's lymphoma and a coding variant in *CASP8* in breast cancer (17,18). But overall, very few candidate studies have yielded convincing results worthy of the enormous investment of time to pursue the biological basis of the association.

Abbreviations: CNV, copy number variation; GWAS, genome-wide association studies; LD, linkage disequilibrium; MAF, minor allele frequency; PSA, prostate serum antigen; SNP, single-nucleotide polymorphism.

In the early part of the new millennium, candidate gene studies expanded in scope, looking at sets of genetic markers across a gene of interest. This transition adopted the use of sets of markers defined on the basis of genetic correlation, known as linkage disequilibrium (LD) discussed below. Often, markers are located in introns or intergenic regions, raising the possibility that genetic variants could alter expression or regulation of a gene, thus not only widening the spectrum of variants to be examined but also increasing the scope of underlying mechanisms. As this approach began to find variants associated with cancer risk, the focus was on markers for risk. For examples, Garcia-Closas *et al.* (19) identified a promising marker near the *VCAM1* gene in association with bladder cancer as part of an exploration of genes in several pathways related to cancer biology. Again, the approach was hypothesis driven, in that specific genes were chosen for the best markers but the scope was enlarging and increasing the number and types of variants explored (20).

In 1996, Risch and Merikangas argued that for complex diseases, such as most cancers, large scale linkage studies will be both difficult and not as well powered to detect susceptibility alleles with low estimated effect sizes, of the type that are probably to contribute in a polygenic model (15,21,22). Instead, they suggested that large-scale association testing could be more efficient and more effective (15,21) in the discovery phase. Moreover, the practicality of collecting large sets of family pedigrees was identified as a daunting, and perhaps overwhelming challenge. Indeed, the age of genome-wide association studies (GWAS) has established the association study as an integral tool for discovering the contribution of common genetic susceptibility alleles to different types of cancer.

The value of conducting statistically sound studies that are well powered has become a central tenet of the GWAS era because of the enormous risk for false-positive discovery. The threshold for discovery has been established at a high level, known as genome-wide significance, which serves two dual purposes (23,24). First, it necessitates careful consideration of the power to detect the effect sizes expected to be observed in the study. Second, the high bar of genome-wide significance protects against the probability of a false-positive finding (25,26). The latter is critical because GWAS are discovery tools that point investigators toward long arduous follow-up studies for unraveling the underlying biology and the pursuit of markers for risk assessment (27).

Background

The scope of genetic variation

Based on the international annotation projects and the sequencing of nearly a dozen full human genomes, the spectrum of human genetic variation is enormous with respect to the types of genetic variation and the magnitude of variants in any given genome (28–34). Although two genomes are estimated to differ by $<0.5\%$, there are at least several million differences, only a small subset of which contributes to disease risk while the majority is probably vestigial. The most common type of variation is a single-nucleotide base substitution, known as the SNP. Next generation sequence analysis has begun to identify the large set of small insertions or deletions in sequence (30,35,36). Progressively, larger structural alterations and copy number variants are fewer in absolute number but impact more bases across the genome (Figure 1).

Most common variants namely those with a minor allele frequency (MAF) $>5\%$ are common to all populations, although the distribution of allele frequencies can vary greatly across the globe (37). Ascertainment estimates for lower frequency variants depend on both the number of subjects as well as the population genetic history of those examined. With next generation sequencing applied to high-profile regions in large numbers, greater complexity in different human populations is emerging, particularly with variants of lower frequency (36,38,39). Interestingly, the scope of structural variants is much greater than previously recognized, though the majority of large-scale polymorphisms appear to be less common, namely <1–5% in unrelated populations, unlike SNPs and insertions and deletions, of which there are millions with frequencies $>5\%$. Accordingly, the GWAS approach in unrelated subjects has been most successfully applied to SNPs and it has been far less successful applied to structural variants, also known as copy number variations (CNVs).

The most common sequence variation in the germ line genome is SNP, which, by definition, is observed in at least 1% of a population. By definition, the MAF is a relative term and applies to the allele with the lower frequency at a locus in a reference population. In many instances, there can be major differences in MAFs between populations with distinct histories. For the common SNPs (MAF $>5\%$), $<10\%$ of SNPs are specific to a given population (28,37). This observation suggests the common ancestry of common SNPs. The literature suggests that there are at least 10 million SNPs with

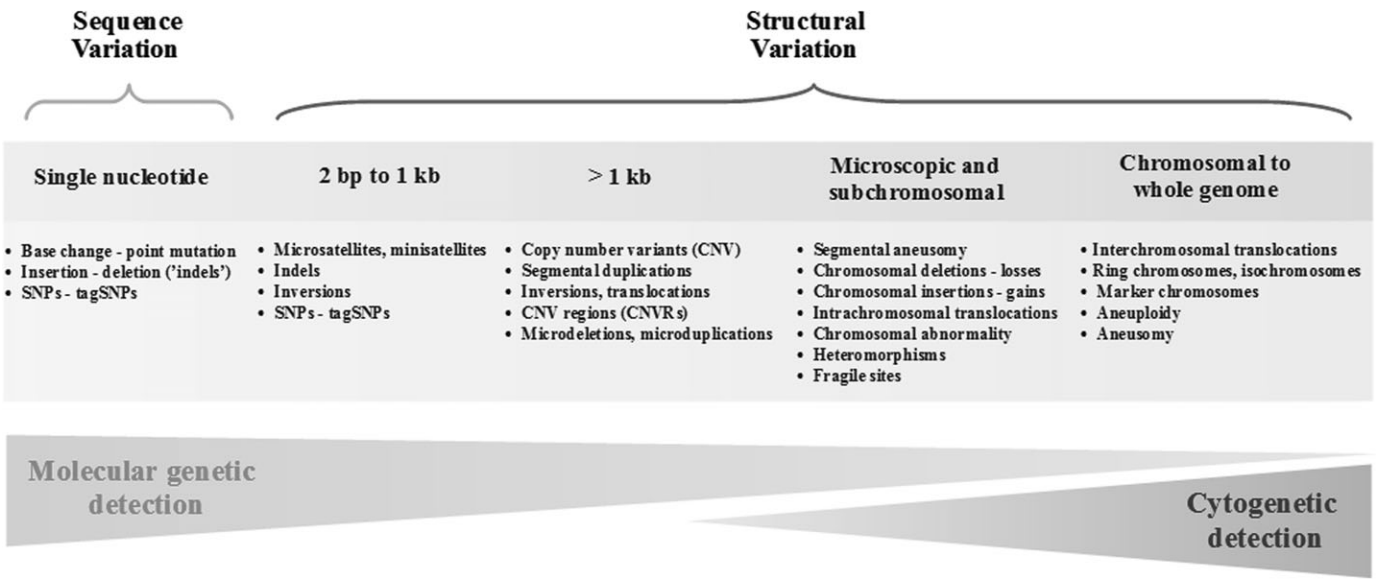

Fig. 1. Types of genetic variations in the human genome. Common types of genetic variations can be categorized into two major groups—those that involve single base changes (e.g. SNPs) and those that alter more than one base (e.g. microsatellites or structural variants).

a MAF >1% (40–42) and 5 million SNPs with a MAF >10% (3,4,40) but recent large-scale sequencing efforts, such as the 1000 Genome project, indicate that these estimates are low (www.1000genomes.org/) (43). In fact, there could be double or triple the earlier estimates. Lastly, there is a small subset of SNPs that are tri-allelic; at a given base on the reference genome, there can be three different bases, though these are rare, they can be formidable technical challenges for quality control metrics.

It is estimated that between 50 000 and 250 000 common SNPs could be biologically active, as non-synonymous coding variants or regulators of gene expression or splicing (7,15). For candidate gene studies, there was a premium assigned to SNPs in coding regions, usually based on *in silico* predictions. These coding SNPs, known as cSNPs, can be divided into non-synonymous variety (which alters the predicted amino acid codon) and synonymous SNPs (which do not alter the codon sequence). The latter are far more common and less probably alter function. Though intense interest has been directed at non-synonymous SNPs, few have been conclusively associated with human diseases and even fewer have corroborative biological data to provide plausibility for the association (7,15). There has been considerable effort to predict the effect of a non-synonymous cSNP and putative conformational protein changes, but the biological significance is based on laboratory evidence only. Recently, it has emerged that there are subset of SNPs that alter regulation or expression of a gene. These regulatory SNPs are difficult to identify using informatic tools and thus have to be defined on the basis of laboratory data (44).

More than 5 million human SNPs of the international public repository for SNPs, known as dbSNP (www.ncbi.nih.gov/SNP/), have been validated to date with genotyping assays by the SNP Consortium and the International HapMap Project (1,28). Until recently, sequence validation was applied to a small subset but this is about to shift with the completion of the 1000 Genome Project, so that the majority of entries will be sequence based (45,46). Historically, many variants in dbSNP are monoallelic, due to either genotyping error or, more probably, sequencing errors (47,48). It is notable that the reported SNPs have been biased toward high-frequency variants in populations of European ancestry. The catalog of uncommon variation, namely SNPs with MAF under 1%, is incomplete but the 1000 Genome Project is expected to generate a catalog of variants between 0.5 and 5% frequency, which will complement the International HapMap of common variants above 5–10%. Already, the latest build of dbSNP has >20 million variants, mainly less common ones. In addition, dbSNP contains downloads from many disease-specific mutation databases, which will make the curation and utility of less common variants even more daunting for analytical approaches toward prioritization of variants for study. Still, the contribution of uncommon variants represents an untapped portion of the genomic architecture and will necessitate new approaches toward mining these variants for cancer susceptibility. Highly penetrant disease mutations are cataloged in a public database, the Online Mendelian Inheritance in Man or OMIM (www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM/).

The spectrum of genetic variation in the genome can range from single base substitutions to small insertions/deletions to structural variations that can be cytologically observed. The short tandem repeat, also known as the microsatellite, represents a class of polymorphisms used in linkage analysis that are defined by repeats of two or more nucleotides but display notable differences in the frequencies of the repeat units. Typically, they are located in non-coding regions. However, most large-scale structural variation is submicroscopic and ranges in size from a few base pairs to thousands of base pairs (49,50). Collectively, the submicroscopic variants are known as CNVs, a focus of intense interest in large-scale association studies. Estimates of segmental duplications in the genome have been suggested to approach 10% of the genome, but most are not common enough to be effectively analyzed using current GWAS (51–53). Current surveys suggest that CNVs are less common than previously reported (54,55) and in fact, perhaps, three-quarters of common CNVs are in LD with common SNPs (55).

Correlation of common genetic variants

It has been observed that the majority of SNPs are not inherited independently but segments on a chromosome, inherited from generation to generation (41,56,57). A central concept in germ line genetics is the inheritance of correlated markers on the same chromosome, known as LD. It is defined as the non-random association between allelic markers on a chromosome and is classically measured using one of two estimators, D′ or r^2 (58). Individual SNPs that are strongly correlated with each other are said to be in LD, but with time and geographic distribution, LD can erode by recombination events (e.g. exchange of genetic material) during meiosis (59).

Haplotypes are defined as sets of SNPs or polymorphisms (e.g. insertions, deletions or large copy events) in strong LD, in which one or more can serve as surrogates for the other markers on the haplotype. A haplotype can be determined in most cases with family trios but in GWAS or large association studies, family structure is usually not available. Still, the offspring haplotype phase can be determined if the parental genotypes are known or established by biochemical methods and then applied to study to best estimate the common haplotypes (58). However, the phasing of haplotypes is more challenging in unrelated subjects but accurate estimates based by well-developed statistical methods that can account for the ambiguity of unobserved haplotypes can provide haplotypes with assigned probabilities (58). Some have argued that haplotypes are preferable for candidate gene studies but for GWAS, the approach is laborious and less nimble in analyzing the thousands of markers genotyped. The methods are not as robust for conducting analysis across thousands of variants.

The appreciation of applying LD to the millions of SNPs observed in human populations that has given rise to the fundamental principle of GWAS, testing across the genome with well-chosen markers that serve as surrogates for untested markers (60–62). The ‘indirect approach’ represents the first step in identifying regions with strong association with cancer or a human trait and relegates the investigation of the optimal variants to study for understanding the biological basis of the association signal (59). The commonly used approach to select optimal SNPs is the ‘greedy algorithm’, which estimates highly correlated SNPs, on the basis of MAFs and creates heuristic bins of ‘tagged’ SNPs. It is the set of tags that function as proxies for the highly correlated untested variants (60).

Practical issues in GWAS

GWAS have emerged as a powerful tool to identify susceptibility loci with low effect sizes in unrelated subjects with specific cancers and related outcomes. Though epidemiologic design is important, in the discovery phase, there has been a relaxation of epidemiologic rigor in order to discover novel regions, mainly because of the need to gather a sufficiently large enough data set to detect low effect sizes. Often, groups have used convenient or publicly available controls for the discovery analysis in GWAS (23), of which the Wellcome Trust Case Control Consortium has been a notable example. These steps could come at a cost, such as a slightly higher rate of false positives, or in related manner, the apparent contradiction of regions or loci that do not robustly replicate in separate scans, suggesting subtle, but real differences related to selection and exposure criteria. Consequently, the estimates are slightly unstable and maybe refined as better studies if analyzed with high quality epidemiologic and environmental exposure data. In order to meet the requirements of a sufficiently large enough data set to observe significant differences between cases and controls, many scans, particularly for rarer cancers, have had to amalgamate data sets.

Replication of results is critical in a separate comparable set of studies (63). The value of replication is to guard against the blizzard of false positives observed with common alleles with low effect sizes. By scaling the studies, GWAS can effectively shed the majority of false positives. The industry standard that has emerged has targeted genome-wide statistical significance for a GWAS with a P value less than between 5×10^{-7} and 1×10^{-8} using either a trend or genotype test, adjusted for minimal cofactors/covariates (23,64–66).

Because GWAS are conducted in unrelated subjects, there has been intense interest in the background population substructure of cases and controls. The capacity to examine thousands of markers with minimal or no LD can be used to effectively discriminate differences in population substructure (67–69). Population stratification is present when there is a measurable difference in the distribution of alleles between subgroups that have different population histories, which can certainly alter association analyses, providing false-positive findings, such as in early case–control studies, in which the cases and controls were drawn from individuals of different populations. Stratification between cases and controls based on differences in exposures can also be problematic, but less so in GWAS. The ability to detect stratification with sets of markers depends on the allele frequencies in each subgroup (70). Subjects with admixture coefficients >15–20% can be removed from association analyses (71) based on attempt to separate subjects into groups and determining the distribution of shared alleles. Further, detection of population stratification is conducted on the GWAS data set to adjust simultaneously for a fixed number of top-ranked principal components resulting from a principal component analysis (67). The search for underlying subgroups in stratified samples can be investigated with genetic markers not linked to the phenotype, using a principal component analysis that yields eigenvectors, used to adjust for possible inflation of test statistics due to stratification (67,72,73).

One of the fundamental reasons for the success of GWAS has been the foresight to collect biospecimens in case–control and cohort studies over the past decades, each of which affords advantages for studying exposures or avoiding survivorship bias. Since the high throughput genotype platforms that analyze thousands of commercially determined SNPs and now CNVs demand high performance DNA, most investigators have used native DNA—either from blood or buccal cells. The latter works quite well when optimally collected and extracted (74). Neither whole genome amplified DNA can be effectively used in GWAS or can materials from tumor tissue (or its adjacent region) due to problems with allelic imbalance. High-quality genotypes are generated using widely accepted quality control metrics for SNP completion, sample completion, heterozygosity scores, testing for fitness for proportion of Hardy–Weinberg equilibrium (70) and assay verification with a second technology (75).

Scanning the genome with SNPs can be performed with commercially available fixed products that provide hundreds of thousands of SNPs, chosen either on the basis of the tag strategy, spacing across the genome or inclusion of obligate SNPs either known or predicted to be functionally important. Great importance has been attached to the extent of 'coverage' afforded by the fixed content chips, which for each commercial product has translated into higher cost for greater coverage (24). The bias of the chips has been to select SNPs that most efficiently tag common SNPs in individuals of European background based on the successive builds of the International HapMap Project (Figure 2). Specifically, the level of coverage is generally measured by determining the percentage of 'bins' tagged by SNPs (with MAF > 5 or 10%) for each of the three HapMap II populations, individuals of European background (known as CEU), Yoruban of West Africa (YRI) and East Asians (CHN and JPN) (24,59,60). Over 500 regions of the genome have now been conclusively associated (e.g. report signals with P value $<5 \times 10^{-7}$) in >100 human diseases or traits (76–78).

The analysis of dense genotyping data can be carried out with publicly available tools in either Genotype Library and Utilities (GLU) or PLINK (79), each of which permits archiving, manipulation and basic analyses of data sets, including assessment of population

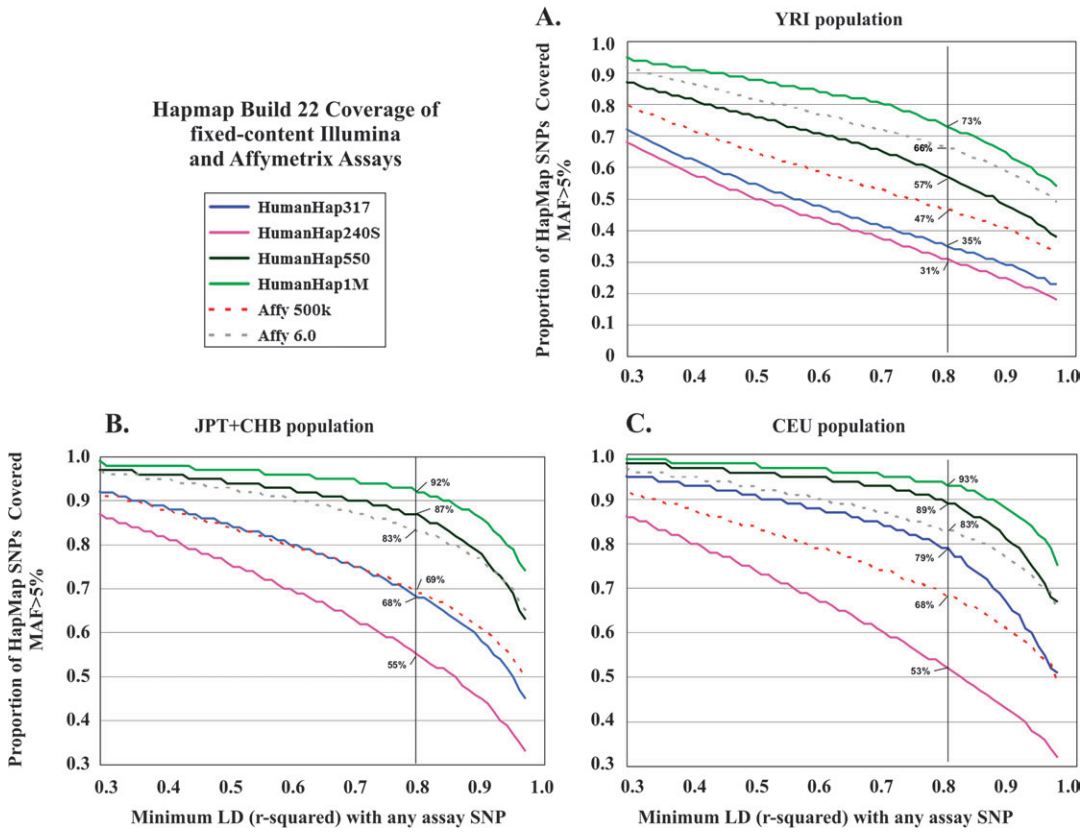

Fig. 2. Coverage of various genotyping platforms on HapMap II SNPs. The coverage of commercially available genotyping platforms in HapMap populations are plotted based on estimates of linkage disequilibrium using r^2, the correlation coefficient. A vertical bar depicts the cut off of an $r^2 = 0.8$, which is commonly used as a threshold to effectively tag monitored SNPs. The three HapMap populations of Phase II are labeled and the percentage estimated at the threshold is provided. (**A**): Coverage plot in Yoruban population (Ibadan, Nigeria), (**B**): coverage plot in Japanese (Tokyo, Japan) and Han Chinese (Bejing, China) and (**C**): coverage plot of US residents with northern and western European ancestry by the Centre d'Etude du Polymorphisme Humain (CEPH).

substructure and association testing for SNPs. CNVs are more challenging because the primary image files have to be analyzed and quality control metrics applied to predict CNVs with varying degrees of probability. It is this latter issue, together with the evolving annotation of CNVs, which has hampered the widespread application of this type of analysis to yield association results comparable to those from common SNPs. Consequently, only a handful of common CNVs have been conclusively associated with complex diseases. In cancer GWAS, only one conclusive finding has been reported, the association of a region on chromosome 1 with the rare pediatric cancer neuroblastoma (80).

The first look at GWAS findings in cancer

Theme and variations

The age of GWAS and cancer have quickly ushered in a new era of discovery of regions that harbor germ line genetic variants (common and uncommon) associated with susceptibility to specific cancers. Currently, >75 regions of the genome (some harboring multiple independent signals) have been conclusively associated with susceptibility to specific cancers. Notably, in a handful of few circumstances, more than one type of cancer maps to the same set of genetic variants but overall, it appears that the contribution of common germ line variation has a strong component of tissue specificity. It is also notable that no single locus identified by the current crop of etiologically driven GWAS has also been shown to influence outcome, as measured by progression, disease stage, metastases or survivorship. This latter observation suggests that the germ line factors responsible for development of a cancer could differ from those genetic factors that sustain carcinogenesis or lead to progression. It is interesting to note that for the 29 independent loci identified in prostate cancer GWAS, so far, not a single locus exclusively associates with the more aggressive form of the disease (65,66,81–84). In the Cancer Genetic Markers of Susceptibility Initiative of a GWAS in prostate cancer, the analysis plan specifically addressed the early and advanced forms of prostate cancer, yet did not identify a locus specific to disease state (65,66,84). Consequently, it will be necessary to conduct distinct GWAS in studies designed to address these important outcomes, but it will most probably require new collections and collaborative networks to achieve the required numbers to discover the low to moderate effect alleles influencing cancer outcomes.

It was unanticipated that GWAS studies in certain cancers would yield many novel regions (e.g. prostate cancer with perhaps 29, breast cancer with 13 and colon with 10) (64,66,75,81–93), whereas other cancers strongly associated with environmental exposures have yielded so few regions: three for lung cancer in primarily smokers and three in bladder cancer despite analysis of sufficiently large data sets. Thus, it is plausible that the effect of tobacco use is substantially stronger than any single region with low estimated effect sizes (below 1.3 in GWAS). The lung cancer findings are also notable in that the strongest signal on chromosome 15q25 maps to a region that has also been identified in GWAS of smoking phenotypes (94–97). Prior to GWAS, it was also considered on the list of candidate genes because it contains nicotine receptors (e.g. *CHNRA3* and *CHRNA5*) (98,99). Further studies are urgently needed in non-smoking cases and controls to discriminate between signals that could be driven by tobacco exposure versus primary carcinogenesis (94). Fine-mapping studies in different populations may accelerate the pinpointing of the set of variants in this region requiring further study to understand the biology underlying the association study.

There are few notable exceptions to the observation that the per allele estimated effect is <1.5 for alleles discovered in cancer GWAS (100). In fact, most are <1.3, and it is anticipated that more will be discovered in the vicinity of 1.1–1.2 as consortial activities permit meta-analyses with larger sets of scanned subjects (Figure 3). Still, it was notable that two recent testicular cancer scans each identified two regions with effect sizes considerably greater than what had been observed previously in cancer GWAS. The loci mapped to regions on chromosomes 5 and 12 that harbored candidate genes previously implicated in testicular development, the ligand for the receptor tyrosine kinase (*KITLG)* and sprouty 4 (*SPRY4*). Moreover, the studies were notable for the high effect sizes detected for chromosome 5, namely >2.5, as well as the biological plausibility of the candidate genes (101,102). This was not surprising in light of the marked increase risk for family members (103,104). Another cancer with a familial aggregation, thyroid cancer, also yielded alleles with relatively high estimated effect sizes, and interestingly, they were detected in a small primary scan (105).

In select GWAS, the findings have pointed to genes previously investigated in that cancer. Pancreatic cancer is a highly lethal disease with a 5-year relative survival of <5% (106), with known risk factors of family history of pancreatic cancer, type 2 diabetes mellitus and cigarette smoking. Interestingly, the first reported GWAS in pancreatic cancer identified a variant in an intron of the ABO blood group antigen, which confirmed a finding suggested 50 years ago (107,108).

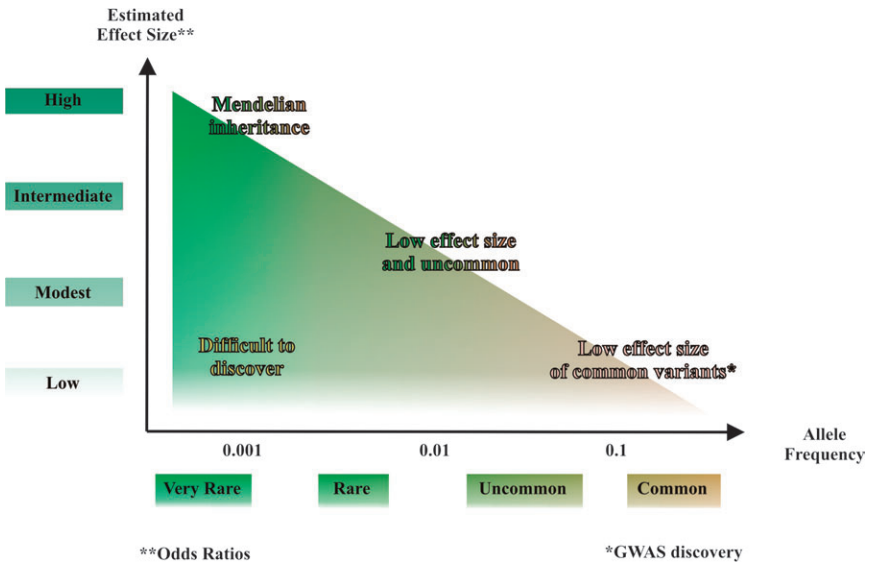

Fig. 3. The relationship between the estimated effect size and the allele frequency of disease susceptibility locus. The majority of disease susceptibility loci identified by GWAS in different cancers have low effect size (per allele estimated effect size of 1.1–1.3).

This is a striking example of how a GWAS hit points to a finding previously described in the epidemiology literature and has been confirmed with a recent study, in which comparable effect sizes have been observed by known blood type (109).

In prostate cancer, the signal on chromosome 10q13 points to a variant in the promoter of the *MSMB* gene, which encodes a protein, PSP94, under intense investigation as a biomarker for prostate cancer (65,89). The T allele of rs10993994, 57 bp centromeric to the first exon of the *MSMB* gene, showed significant association with prostate cancer in two independent studies (65,89), and it is known to have influence in the *MSMB* gene expression (prostate secretory protein 94, PSP94) in tumor (110,111). Now that the region has been extensively resequenced, further investigation of additional variants in strong LD with rs10993994 is warranted and it is possible that a neighboring gene, *NCOA4*, could also be a candidate gene for analysis because it is an androgen receptor coactivator.

A GWAS of neuroblastoma, a rare pediatric cancer, has implicated three different chromosomal regions, one of which is a copy number variation at chromosome 1q21.1 (80,112,113). The first region is at 6p22 and it is plausible that the risk alleles have dosage effect on the severity of disease by subgrouping patients into patients of metastatic stage 4, patients with somatic *MYCN* amplification and patients with relapse. The second region is at 2q35 within the *BARD1* gene (112).

Despite the enormous effort focused on choosing candidate genes or pathways, based on current models, so far, the results of cancer GWAS have pointed to primarily new or unknown regions and genes. However, there are a few notable exceptions, such as two GWAS of pediatric lymphoblastic leukemia, which have uncovered three sets of markers pointing to genes involved in B-cell development (114,115), but the clustering of related genes has not been observed. Moreover, for a disease such as breast cancer, which has been epidemiologically linked to hormones, surprisingly, none of the major signals map to regions harboring estrogen/progesterone genes in women of European background. However, in a scan of Asian women, a GWAS convincingly discovered markers near the estrogen receptor alpha (known as *ESR1*) (93).

Discovering more complexity

GWAS have uncovered a series of possible interesting and unexpected relationships between different diseases. For example, three of the regions identified in prostate cancer GWAS also map to type two diabetes susceptibility regions. For some time, there has been a controversial literature reporting an inverse relationship between type two diabetes and prostate cancer; it is further speculated that the protection against prostate cancer is more apparent several years after the diagnosis of diabetes. For two of regions, the markers appear to be inversely related, namely the apparent risk allele for prostate cancer is protective for diabetes for *HNF1B* on chromosome 17q24 and for *THADA* on chromosome 2p21. The signal on chromosome 7p15 localizes to intron 2 of *JAZF1*, a very large gene, whereas the diabetes signal, as well SNPs for height, body stature and systemic lupus erythematosus are localized to a distinct region >200 kb away in intron 1 with no residual LD, suggesting different variants.

Differences in study design can lead to important observations related to both the genetic and environmental contributions to cancer etiology. In one notable instance, two distinct GWAS efforts in prostate cancer have yielded different results for a region of chromosome, 19q13.33, that harbors the gene responsible for the prostate serum antigen (PSA), used by many, but not all for screening for prostate cancer (116,117). In one study, that used clinically advanced cases with controls that had low PSA levels, a strong signal for a SNP in *KLK3* was observed, replicating with a substantially lower degree of statistical significance in the follow-up studies, whereas in Cancer Genetic Markers of Susceptibility Initiative, comprised of mainly cohort studies, there was little effect for prostate cancer risk (39,89,118,119). In fact, the Cancer Genetic Markers of Susceptibility Initiative analysis reported that the SNP in the region of *KLK3* was associated with PSA levels, raising the possibility that the locus could be related to PSA levels instead of prostate carcinogenesis, though it is possible it could be a both but further studies are needed. Indeed, now that the *KLK3* region has been resequenced, it will be possible to investigate this issue with the optimal markers (36).

Most studies have relied on combining data from different designs and often combining histologic or molecular subtypes of a classically defined cancer. The result has been to identify regions that appear to be associated with biological processes common to the development of a tissue-specific type of cancer. For example, the follow-up analysis of the initial set of signals identified in breast cancer GWAS suggests that there could be a differential effect for some regions based on estrogen receptor status for some regions (120). The preponderance of estrogen receptor-positive cases in the discovery studies certainly could have contributed to this observation, but additional reports have identified regions with stronger effects in estrogen receptor-positive subjects (92). In other GWAS, subtype GWAS have yielded convincing findings for a histologic subtype, such as the chromosome 5p15.33 locus in lung cancer (in predominately smokers), which is significantly associated in the adenocarcinoma subtype but not in squamous cell carcinoma (121,122). Similarly, in non-Hodgkin's lymphoma, distinct regions have been identified in the chronic lymphocytic leukemia (114) and follicular subtypes (123). On the other hand, for the associations with high effect sizes in testicular cancer, there was no appreciable difference by subtype analysis for seminoma and nonseminoma cancers, suggesting the common contribution of the two regions to testicular carcinogenesis (101,102,124).

Based on follow-up fine mapping of the regions, often using HapMap chosen SNPs or those defined by comprehensive resequence analysis (36,38,39), intense effort has focused on the investigation of the genomic architecture of each GWAS region. It is plausible that more than one common variant, each with small effect sizes, could contribute to cancer susceptibility and in fact, this has been demonstrated in three regions identified in prostate cancer susceptibility. For 8q24, there are at least four distinct prostate cancer susceptibility loci in men of European background (66,82,84,85,90,125). In men of other backgrounds (e.g. African, East Asian or Latino/admixed), it is possible that even more population-specific loci could be important and perhaps partially explain some of the disease disparity among different ethnic groups (85,90). For the *HNF1B* locus on chromosome 17q24, further mapping identified a second independent signal (126). Similarly, the gene desert of 11q13 harbors at least two independent signals and perhaps more (127).

Cancer GWAS Nexus regions

8q24, a cancer susceptibility region for many unrelated cancers

A region of ~600 kb, centromeric to the well studied, *MYC* oncogene, is a region that has been repeatedly discovered to harbor distinct independent markers associated with cancer risk (Figure 4). *MYC* encodes for nuclear phosphoprotein that involves in growth regulation, cell differentiation and apoptosis, and its amplification/overexpression is a frequent event in bladder tumors (128,129). The findings have unexpectedly found that prostate, breast, colorectal, bladder and perhaps ovarian cancers are associated with common genetic variants in this region (66,75,82,88,90,130–134). The region is also notable because it is frequently amplified in epithelial cancers and does not harbor candidate genes, but instead several pseudogenes, whose function and presence are not well established. In this regard, the findings of 8q24 attest to the complexity of the region and the likelihood that regulatory elements of both *MYC* and other regions could underlie the cancer susceptibility.

The 8q24 region was first implicated as a prostate cancer risk locus by a genome-wide linkage scan in Icelandic men, followed by identification of an allele of the microsatellite marker, DG8S737, and A allele of rs1447295 from replication association studies in three case–control samples of European ancestry from Iceland, Sweden and USA (125). The region was also discovered by an admixture mapping

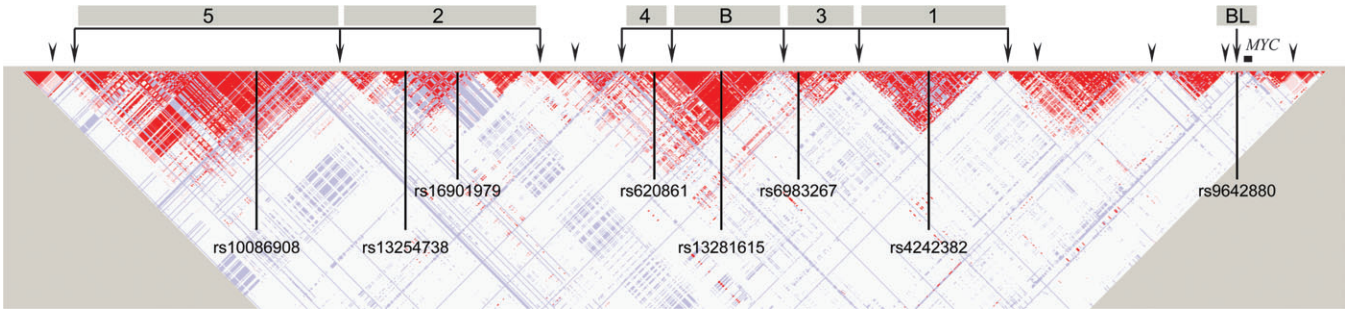

Fig. 4. Linkage disequilibrium pattern and cancer susceptibility loci indentified in 8q24 region. The 8q24 region harbors multiple cancer susceptibility loci identified by GWAS. The linkage disequilibrium heat map was drawn using HapMap I + II release 22 CEU data from 127 948 to 128 950 kb genomic region (reference build 36.3). The arrowheads indicate probable recombination hotspots according to the HapMap I + II. Five distinct regions have been associated with prostate cancer risk (regions 1–5). Region 3 is also conclusively associated with colorectal cancer and precancerous colorectal adenomas. Region B harbors a breast cancer susceptibility locus rs13281615, and BL indicate a bladder cancer susceptibility locus rs9642880, which is telomeric to the region 1, and ∼30 kb centromeric to the *MYC* oncogene.

in African-Americans (135). The SNP, rs1447295, was reconfirmed by a large nested case–control study using 6637 cases and 7361 matched controls (91). Independent of the rs1447295, which marked as 'region 1', two independent loci, rs16901979 and rs6983267, marked as region 2 and region 3, respectively, centromeric to the region 1 were identified by three independent studies (66,82,90). Notably, the rs16901979 showed clear association in African-Americans with higher risk allele frequency than Europeans. In two recent studies, another independent prostate cancer susceptibility locus rs620861 was identified, located in between region 2 and region 3 and overlapping with a region previously identified in a breast cancer GWAS (81,84,136).

For colorectal cancer, four different studies reported the same variant, rs6983267 (in region 3 of prostate cancer), as the strongest signal by GWAS (88,90,132,137). Recently, published work has begun to generate insights in the functional nature of the rs6983267 variant, which has only two other variants in strong LD compared with rsw1447295 with 49 variants in strong LD (36,138,139). The two studies suggest that in colorectal cancer, rs6983267 shows long-range interaction with MYC as well as possible enhancement of the Wnt-signaling pathway. Interestingly, the prostate specific effect is more complex and as of now, not well explained except for the presence of multiple regions across the 600 kb of 8q24.

Kiemeney *et al.* (130) reported that the T allele of rs9642880 located ∼30 kb upstream of *MYC* oncogene showed significant association with bladder cancer (odds ratio = 1.22, $P = 9.34 \times 10^{-12}$). Wu *et al.* (140) reported that rs2294008 located in exon 1 of *PSCA* on the other side of *MYC* is significantly associated with bladder cancer risk. Since the SNP, rs2294008, is located in the exon 1 of *PSCA* and yields a missense variant that alters the start codon, Wu *et al.* further performed an *in vitro* reporter assay using the four most frequent haplotypes of the *PSCA* 5′ upstream region including rs2294008 and showed significantly lower promoter activity of the T allele-containing haplotypes.

5p15.33

Common variants in the *TERT-CLPTM1L* locus on 5p15.33 have been identified by GWAS to harbor susceptibility alleles for cancer of the brain and lung (96,97,122,141,142). For lung cancer, it appears that the signal is strongly associated with the adenocarcinoma subtype and not squamous or other subtypes (122). In the region, there is an attractive candidate gene, *TERT*, the reverse transcriptase component of the telomerase a gene that is critical for telomere replication and stabilization by controlling telomere length. TERT promotes epithelial proliferation and telomere maintenance has been implicated in the progression from *KRAS*-activated adenoma to adenocarcinoma in a murine model (143,144). There is additional evidence for associations with cancer of the bladder, prostate, uterine cervix and skin including basal cell carcinoma and melanoma based on candidate studies in follow-up of GWAS hits (145).

This region is particularly interesting because of the scope and spectrum of allele frequencies associated with diseases. Mutations in the *TERT* gene have been described in acute myelogenous leukemia and in the inherited bone marrow failure family pedigrees with dyskeratosis congenita, a cancer predisposition syndromes (146,147). Mutations in the *TERT* gene have also been described in patients with idiopathic pulmonary fibrosis (148,149) and in families with hematologic disorders and serious liver fibrosis (150). Mutations in *TERT* have also been shown to result in shorter telomeres and explain a subset of those with familial idiopathic pulmonary fibrosis (151).

Conclusions

The age of genome-wide association studies in cancer have ushered in a new era of discovery of regions of the genome harboring common genetic susceptibility alleles that require extensive effort to map the signal to define the optimal variants for investigating the biological basis of the association. For nearly all signals identified, the markers have not immediately uncovered variants that can easily explain the signal and in most cases, appear to be variants not in coding regions that instead of shifting the amino acid sequence, probably alter the regulation of one or more complex genetic processes. In this regard, GWAS are the first step toward identifying novel regions and pathways associated with both primary carcinogenesis and probably gene–environment interactions.

To make sense of the known GWAS signals and to find more signals, some that could explain major disparities in incidence and outcomes by ethnic backgrounds, it will be critical to conduct GWAS in populations with distinct population genetic histories (and different underlying LD structures) as well as to map known hits in other populations. The age of GWAS has not only uncovered new regions but perhaps provided insights in a subset of the regions that require refined analyses, such as the effect of tobaccos usage and lung cancer risk to unravel the complex nature of these types of cancer.

The recent genomic revolution has produced a comprehensive map of genetic variation that has enabled research to scan the genome in search of statistically sound signals worthy of follow-up. However, the ability to survey environmental and lifestyle exposures is not nearly as advanced, thus hampering the opportunity to explore the dynamic relationship between genomic variants and the environment. Lastly, the age of GWAS is actually the beginning of a new age, one characterized by many new regions of the genome worthy of pursuit as candidate genes to explore the common as well as uncommon variants that contribute to the risk of different cancers.

Acknowledgements

Conflict of Interest Statement: None declared.

References

1. The International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426,** 789–796.
2. Collins,F.S. *et al*. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.
3. Lander,E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
4. Venter,J.C. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
5. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437,** 1299–1320.
6. Sachidanandam,R. *et al*. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
7. Chanock,S. (2001) Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease. *Dis. Markers*, **17**, 89–98.
8. Miki,Y. *et al*. (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, **266**, 66–71.
9. NIH/CEPH Collaborative Mapping Group. (1992) A comprehensive genetic linkage map of the human genome. NIH/CEPH Collaborative Mapping Group. *Science*, **258**, 67–86.
10. Elston,R.C. *et al*. (2001) Overview of model-free methods for linkage analysis. *Adv. Genet.*, **42**, 135–150.
11. Malkin,D. *et al*. (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, **250**, 1233–1238.
12. Hussussian,C.J. *et al*. (1994) Germline p16 mutations in familial melanoma. *Nat. Genet.*, **8**, 15–21.
13. Kamb,A. *et al*. (1994) Analysis of the p16 gene (CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. *Nat. Genet.*, **8**, 23–26.
14. Wooster,R. *et al*. (1995) Identification of the breast cancer susceptibility gene BRCA2. *Nature*, **378**, 789–792.
15. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
16. Garcia-Closas,M. *et al*. (2005) NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet*, **366**, 649–659.
17. Rothman,N. *et al*. (2006) Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. *Lancet Oncol.*, **7**, 27–38.
18. Cox,A. *et al*. (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.*, **39**, 352–358.
19. Garcia-Closas,M. *et al*. (2007) Large-scale evaluation of candidate genes identifies associations between VEGF polymorphisms and bladder cancer risk. *PLoS Genet.*, **3**, e29.
20. Dunning,A.M. *et al*. (2009) Association of ESR1 gene tagging SNPs with breast cancer risk. *Hum. Mol. Genet.*, **18**, 1131–1139.
21. Risch,N. (2001) The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol. Biomarkers Prev.*, **10**, 733–741.
22. Risch,N. *et al*. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
23. The Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447,** 661–678.
24. Barrett,J.C. *et al*. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.
25. O'Berg,M.T. (1980) Epidemiologic study of workers exposed to acrylonitrile. *J. Occup. Med.*, **22**, 245–252.
26. Wolff,M.S. *et al*. (1993) Blood levels of organochlorine residues and risk of breast cancer. *J. Natl. Cancer Inst.*, **85**, 648–652.
27. Erichsen,H.C. *et al*. (2004) SNPs in cancer research and treatment. *Br. J. Cancer*, **90**, 747–751.
28. Frazer,K.A. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
29. Kidd,J.M. *et al*. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature*, **453**, 56–64.
30. Levy,S. *et al*. (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
31. Ng,S.B. *et al*. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
32. Wang,J. *et al*. (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.
33. Wheeler,D.A. *et al*. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.
34. Kim,J.I. *et al*. (2009) A highly annotated whole-genome sequence of a Korean individual. *Nature*, **460**, 1011–1015.
35. Harismendy,O. *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32.
36. Yeager,M. *et al*. (2008) Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum. Genet.*, **124**, 161–170.
37. Hinds,D.A. *et al*. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
38. Yeager,M. *et al*. (2009) Comprehensive resequence analysis of a 97 kb region of chromosome 10q11.2 containing the MSMB gene associated with prostate cancer. *Hum. Genet.*, **126**, 743–750.
39. Parikh,H. *et al*. (2009) A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33. *Hum. Genet.,* in press.
40. Kruglyak,L. *et al*. (2001) Variation is the spice of life. *Nat. Genet.*, **27**, 234–236.
41. Reich,D.E. *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
42. Reich,D.E. *et al*. (2003) Quality and completeness of SNP databases. *Nat. Genet.*, **33**, 457–458.
43. Hayden,E.C. (2008) International genome project launched. *Nature*, **451**, 378–379.
44. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nat. Genet.*, **33**, 439–440.
45. Packer,B.R. *et al*. (2006) SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes. *Nucleic Acids Res.*, **34**, D617–D621.
46. Stephens,M. *et al*. (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. *Nat. Genet.*, **38**, 375–381.
47. Marth,G. *et al*. (2003) Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl Acad. Sci. USA*, **100**, 376–381.
48. Marth,G.T. *et al*. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.
49. McCarroll,S.A. *et al*. (2007) Copy-number variation and association studies of human disease. *Nat. Genet.*, **39**, S37–S42.
50. Scherer,S.W. *et al*. (2007) Challenges and standards in integrating surveys of structural variation. *Nat. Genet.*, **39**, S7–S15.
51. Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.
52. Bailey,J.A. *et al*. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
53. Bailey,J.A. *et al*. (2002) Recent segmental duplications in the human genome. *Science*, **297**, 1003–1007.
54. Buckley,P.G. *et al*. (2005) Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet.*, **21**, 315–317.
55. McCarroll,S.A. *et al*. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, **40**, 1166–1174.
56. Bonnen,P.E. *et al*. (2002) Haplotype and linkage disequilibrium architecture for human cancer-associated genes. *Genome Res.*, **12**, 1846–1853.
57. Sabeti,P.C. *et al*. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
58. Slatkin,M. (2008) Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, **9**, 477–485.
59. Orr,N. *et al*. (2008) Common genetic variation and human disease. *Adv. Genet.*, **62**, 1–32.
60. Carlson,C.S. *et al*. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
61. Cardon,L.R. *et al*. (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet.*, **19**, 135–140.
62. Johnson,G.C. *et al*. (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
63. Chanock,S.J. *et al*. (2007) Replicating genotype-phenotype associations. *Nature*, **447**, 655–660.
64. Thomas,G. *et al*. (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.*, **41**, 579–584.

65. Thomas,G. *et al*. (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, **40**, 310–315.
66. Yeager,M. *et al*. (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645–649.
67. Yu,K. *et al*. (2008) Population substructure and control selection in genome-wide association studies. *PLoS One*, **3**, e2551.
68. Patterson,N. *et al*. (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.
69. Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
70. Ryckman,K. *et al*. (2008) Calculation and use of the Hardy-Weinberg model in association studies. *Curr. Protoc. Hum. Genet.*, **57**, 1.18.1–1.18.11.
71. Falush,D. *et al*. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.
72. Devlin,B. *et al*. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.
73. Pritchard,J.K. *et al*. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, **65**, 220–228.
74. Feigelson,H.S. *et al*. (2007) Successful genome-wide scan in paired blood and buccal samples. *Cancer Epidemiol. Biomarkers Prev.*, **16**, 1023–1025.
75. Easton,D.F. *et al*. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.
76. Hindorff,L.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
77. Manolio,T.A. *et al*. (2008) A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.*, **118**, 1590–1605.
78. Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
79. Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
80. Diskin,S.J. *et al*. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. *Nature*, **459**, 987–991.
81. Gudmundsson,J. *et al*. (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1122–1126.
82. Gudmundsson,J. *et al*. (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.*, **39**, 631–637.
83. Gudmundsson,J. *et al*. (2008) Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nat. Genet.*, **40**, 281–283.
84. Yeager,M. *et al*. (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **41**, 1055–1057.
85. Eeles,R.A. *et al*. (2009) Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.*, **41**, 1116–1121.
86. Houlston,R.S. *et al*. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet*, **40**, 1426–1435.
87. Hunter,D.J. *et al*. (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.
88. Zanke,B.W. *et al*. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **39**, 989–994.
89. Eeles,R.A. *et al*. (2008) Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.*, **40**, 316–321.
90. Haiman,C.A. *et al*. (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.*, **39**, 638–644.
91. Schumacher,F.R. *et al*. (2007) A common 8q24 variant in prostate and breast cancer from a large nested case-control study. *Cancer Res.*, **67**, 2951–2956.
92. Stacey,S.N. *et al*. (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **39**, 865–869.
93. Zheng,W. *et al*. (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.*, **41**, 324–328.
94. Chanock,S.J. *et al*. (2008) Genomics: when the smoke clears. *Nature*, **452**, 537–538.
95. Hung,R.J. *et al*. (2008) A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*, **452**, 633–637.
96. McKay,J.D. *et al*. (2008) Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.*, **40**, 1404–1406.
97. Wang,Y. *et al*. (2008) Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat. Genet.*, **40**, 1407–1409.
98. Bierut,L.J. *et al*. (2007) Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.*, **16**, 24–35.
99. Caporaso,N. *et al*. (2009) Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One*, **4**, e4653.
100. Easton,D.F. *et al*. (2008) Genome-wide association studies in cancer. *Hum. Mol. Genet.*, **17**, R109–R115.
101. Kanetsky,P.A. *et al*. (2009) Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nat. Genet.*, **41**, 811–815.
102. Rapley,E.A. *et al*. (2009) A genome-wide association study of testicular germ cell tumor. *Nat. Genet.*, **41**, 807–810.
103. Skinner,D.G. (1983) *Urological Cancer.* Grune & Stratton, New York.
104. Swerdlow,A.J. *et al*. (1997) Risks of breast and testicular cancers in young adult twins in England and Wales: evidence on prenatal and genetic aetiology. *Lancet*, **350**, 1723–1728.
105. Gudmundsson,J. *et al*. (2009) Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations. *Nat. Genet.*, **41**, 460–464.
106. Jemal,A. *et al*. (2008) Cancer statistics, 2008. *CA Cancer J. Clin.*, **58**, 71–96.
107. Amundadottir,L. *et al*. (2009) Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.*, **41**, 986–990.
108. Bodmer,W. *et al*. (2008) Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.*, **40**, 695–701.
109. Wolpin,B.M. *et al*. (2009) ABO blood group and the risk of pancreatic cancer. *J. Natl Cancer Inst.*, **101**, 424–431.
110. Chang,B.L. *et al*. (2009) Fine mapping association study and functional analysis implicate a SNP in MSMB at 10q11 as a causal variant for prostate cancer risk. *Hum. Mol. Genet.*, **18**, 1368–1375.
111. Liu,P. *et al*. (2008) Familial aggregation of common sequence variants on 15q24-25.1 in lung cancer. *J. Natl Cancer Inst.*, **100**, 1326–1330.
112. Capasso,M. *et al*. (2009) Common variations in BARD1 influence susceptibility to high-risk neuroblastoma. *Nat. Genet.*, **41**, 718–723.
113. Maris,J.M. *et al*. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.*, **358**, 2585–2593.
114. Papaemmanuil,E. *et al*. (2009) Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat. Genet.*, **41**, 1006–1010.
115. Trevino,L.R. *et al*. (2009) Germline genomic variants associated with childhood acute lymphoblastic leukemia. *Nat. Genet.*, **41**, 1001–1005.
116. Andriole,G.L. *et al*. (2009) Mortality results from a randomized prostate-cancer screening trial. *N. Engl. J. Med.*, **360**, 1310–1319.
117. Schroder,F.H. *et al*. (2009) Screening and prostate-cancer mortality in a randomized European study. *N. Engl. J. Med.*, **360**, 1320–1328.
118. Ahn,J. *et al*. (2008) Variation in KLK genes, prostate-specific antigen and risk of prostate cancer. *Nat. Genet.*, **40**, 1032–1034; author reply 1035–1036.
119. Eeles,R. *et al*. (2008) Reply to ''Variation in KLK genes, prostate-specific antigen and risk of prostate cancer''. *Nat. Genet.*, **40**, 1035–1036.
120. Garcia-Closas,M. *et al*. (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.*, **4**, e1000054.
121. Broderick,P. *et al*. (2009) Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res.*, **69**, 6633–6641.
122. Landi,M.T. *et al*. (2009) A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.*, **85**, 679–691.
123. Skibola,C.F. *et al*. (2009) Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat. Genet.*, **41**, 873–875.
124. Chanock,S. (2009) High marks for GWAS. *Nat. Genet.*, **41**, 765–766.
125. Amundadottir,L.T. *et al*. (2006) A common variant associated with prostate cancer in European and African populations. *Nat. Genet.*, **38**, 652–658.
126. Sun,J. *et al*. (2008) Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. *Nat. Genet.*, **40**, 1153–1155.
127. Zheng,S.L. *et al*. (2009) Two independent prostate cancer risk-associated loci at 11q13. *Cancer Epidemiol. Biomarkers Prev.*, **18**, 1815–1820.
128. DePinho,R.A. *et al*. (1991) myc family oncogenes in the development of normal and neoplastic cells. *Adv. Cancer Res.*, **57**, 1–46.

129. Mhawech-Fauceglia,P. *et al*. (2006) Genetic alterations in urothelial bladder carcinoma: an updated review. *Cancer*, **106**, 1205–1216.
130. Kiemeney,L.A. *et al*. (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.*, **40**, 1307–1312.
131. Ghoussaini,M. *et al*. (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.*, **100**, 962–966.
132. Tomlinson,I. *et al*. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.*, **39**, 984–988.
133. Tomlinson,I.P. *et al*. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.*, **40**, 623–630.
134. Tenesa,A. *et al*. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.*, **40**, 631–637.
135. Freedman,M.L. *et al*. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA*, **103**, 14068–14073.
136. Al Olama,A.A. *et al*. (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1058–1060.
137. Gruber,S.B. *et al*. (2007) Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol. Ther.*, **6**, 1143–1147.
138. Tuupanen,S. *et al*. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.*, **41**, 885–890.
139. Pomerantz,M.M. *et al*. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.*, **41**, 882–884.
140. Wu,X. *et al*. (2009) Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. *Nat. Genet.*, **41**, 991–995.
141. Wrensch,M. *et al*. (2009) Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat. Genet.*, **41**, 905–908.
142. Shete,S. *et al*. (2009) Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.*, **41**, 899–904.
143. Choi,J. *et al*. (2008) TERT promotes epithelial proliferation through transcriptional control of a Myc- and Wnt-related developmental program. *PLoS Genet.*, **4**, e10.
144. Sweet-Cordero,A. *et al*. (2006) Comparison of gene expression and DNA copy number changes in a murine model of lung cancer. *Genes Chromosomes Cancer*, **45**, 338–348.
145. Rafnar,T. *et al*. (2009) Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat. Genet.*, **41**, 221–227.
146. Calado,R.T. *et al*. (2009) Constitutional hypomorphic telomerase mutations in patients with acute myeloid leukemia. *Proc. Natl Acad. Sci. USA*, **106**, 1187–1192.
147. Yamaguchi,H. *et al*. (2005) Mutations in TERT, the gene for telomerase reverse transcriptase, in aplastic anemia. *N. Engl. J. Med.*, **352**, 1413–1424.
148. Tsakiri,K.D. *et al*. (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc. Natl Acad. Sci. USA*, **104**, 7552–7557.
149. Mushiroda,T. *et al*. (2008) A genome-wide association study identifies an association of a common variant in TERT with susceptibility to idiopathic pulmonary fibrosis. *J. Med. Genet.*, **45**, 654–656.
150. Calado,R.T. *et al*. (2009) A spectrum of severe familial liver disorders associate with telomerase mutations. *PLoS ONE*, **4**, e7926.
151. Armanios,M.Y. *et al*. (2007) Telomerase mutations in families with idiopathic pulmonary fibrosis. *N. Engl. J. Med.*, **356**, 1317–1326.

Received October 30, 2009; revised October 30, 2009;
accepted October 30, 2009